



The Price of Local Fairness in Multistage Selection

Vitalii Emelianov, George Arvanitakis, Nicolas Gast, Krishna P Gummadi,
Patrick Loiseau

► To cite this version:

Vitalii Emelianov, George Arvanitakis, Nicolas Gast, Krishna P Gummadi, Patrick Loiseau. The Price of Local Fairness in Multistage Selection. IJCAI-2019 - Twenty-Eighth International Joint Conference on Artificial Intelligence, Aug 2019, Macao, France. pp.5836-5842, 10.24963/ijcai.2019/809 . hal-02145071

HAL Id: hal-02145071

<https://inria.hal.science/hal-02145071>

Submitted on 18 Jun 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The Price of Local Fairness in Multistage Selection

Vitalii Emelianov¹, George Arvanitakis¹, Nicolas Gast¹, Krishna Gummadi² and Patrick Loiseau^{1,2}

¹Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG

²Max Planck Institute for Software Systems

{vitalii.emelianov, george.arvanitakis, nicolas.gast, patrick.loiseau}@inria.fr, gummadi@mpi-sws.org

Abstract

The rise of algorithmic decision making led to active researches on how to define and guarantee fairness, mostly focusing on one-shot decision making. In several important applications such as hiring, however, decisions are made in multiple stage with additional information at each stage. In such cases, fairness issues remain poorly understood.

In this paper we study fairness in k -stage selection problems where additional features are observed at every stage. We first introduce two fairness notions, local (per stage) and global (final stage) fairness, that extend the classical fairness notions to the k -stage setting. We propose a simple model based on a probabilistic formulation and show that the locally and globally fair selections that maximize precision can be computed via a linear program. We then define the *price of local fairness* to measure the loss of precision induced by local constraints; and investigate theoretically and empirically this quantity. In particular, our experiments show that the price of local fairness is generally smaller when the sensitive attribute is observed at the first stage; but globally fair selections are more locally fair when the sensitive attribute is observed at the second stage—hence in both cases it is often possible to have a selection that has a small price of local fairness and is close to locally fair.

1 Introduction

The rise of algorithmic decision making in applications ranging from hiring to crime prediction [Perry *et al.*, 2013] has raised critical concerns regarding potential unfairness towards groups with certain traits, supported by recent empirical evidences of discrimination [Lambrecht and Tucker, 2018; Larson *et al.*, 2016]. This led to a fast-growing body of literature on what fairness in algorithmic decision making is and how to guarantee it (see related works below).

The existing literature typically considers one-shot decision processes whereby, from a set of features observed about an individual—one of them being a ‘sensitive feature’ based on which discrimination is defined—, one needs to decide

whether or not to “select” him/her (where select can mean hire, grant a loan or parole, etc. depending on the context). The problem in this setting is how to learn a decision rule from past data that respects certain fairness constraints. In many applications, however, decisions are made in multiple stages. In hiring for instance, a subset of candidates is first selected for interview based on resume (or high-level candidate’s information) and a final selection is then made from the subset of interviewed candidates. In police practices, there are often multiple stages of decisions with increasingly high levels of investigation of the individuals not released at the previous stage; as for instance in the famous stop-question-and-frisk practice by the New-York City Police Department.

A distinctive specificity of the multistage setting, besides the fact that decisions are made in multiple stages, is that in many cases additional features get known at later stages for the subset of individuals selected at earlier stages, but one needs to make the early-stage selection without observing those features. This raises a number of new questions that are fundamental to fair multistage selection. First, given that there are multiple layers of decisions, *how should fairness be defined?* In particular, should it be defined at each individual stage, on the final decision, or otherwise? Second, given that one has to make decisions with only partial information at early stages, *how to make an optimal selection?* Finally, given that the sensitive feature can be observed at different stages, *is it better to observe the sensitive feature at earlier or later stages (for both fairness and utility)?* This last question intuitively relates to recurrent public debates such as “should gender identification be removed from CVs?”.

In this paper, we study the k -stage selection problem, in which there is a fixed limit (or budget) of candidates that can be selected at each stage (as is natural in the applications discussed). To tackle the questions above, we propose a simple model based on a probabilistic formulation in which we assume perfect knowledge of the joint distribution of features at all stages and of the conditional probability of being a desirable candidate conditioned on feature values. Based on this model, we are then able to make the following contributions.

We introduce two meaningful notions of fairness for the k -stage setting: *local fairness* (the selection is fair at each stage) and *global fairness* (only the final selection needs to be fair). These definitions extend classical group fairness notions for one-stage decision making (such as demographic parity or

equal opportunity) to the multistage setting and they apply regardless of when the sensitive feature is observed (at first stage or later). We show that local fairness implies global fairness and we propose a linear formulation of the problem that allows us to compute the selection algorithm that maximizes precision while satisfying (local or global) fairness and per-stage budget constraints in expectation. As local fairness is a more restrictive condition, the precision of the optimal globally fair algorithm is naturally higher than for the locally fair algorithm. To capture this gap, we define the *price of local fairness* ($PoLF$) as the ratio of the two and prove a simple upper bound—showing that imposing local fairness cannot be arbitrarily bad. We also define the notion of violation of local fairness ($VoLF$) to capture how far from locally fair the optimal globally fair algorithm is.

Finally, we conduct a numerical study in a two-stage setting using three classical datasets. Our results show that the $PoLF$ can be large (up to 1.6 in some cases). This implies that in some cases, enforcing local fairness constraints can reduce the precision by 60% compared to a globally fair algorithm. The $VoLF$ is also sometimes large (up to 0.6 in our experiments), which means that imposing only a global fairness constraint can be highly unfair at intermediate stages. We finally compare what happens when the sensitive feature is observed at the first stage or at the second stage. We find that the $PoLF$ is generally higher when the sensitive feature is observed at the second stage; while conversely the $VoLF$ is generally higher when the sensitive feature is observed at the first stage. These results show that, in most cases, it is possible to get at least approximate fairness at each stage and precision close to globally-fair optimal together; either by imposing local fairness if the sensitive feature is observed at first stage (where $PoLF$ is small) or by hiding the sensitive feature at first stage and using a globally fair algorithm (which is close to locally fair since $VoLF$ is then small).

Overall, our results provide intuitive answers towards better understanding fairness in multistage selection. To that end, we intentionally used the simplest model that captures the main features of a multistage selection problem and how an optimal selection algorithm is affected by the fairness notion considered and the time at which the sensitive feature is observed—rather than using a more practical but complex model. We believe that it is a good abstraction to start with, but we elaborate further on our model’s limitations in Section 6. Some details (proofs, additional formalization and experimental results) are omitted throughout the paper and deferred to the appendices.

Related Works

As mentioned earlier, there have been many recent works on defining fairness and constructing algorithms that respect those definitions for the case of one-stage decision making [Pedreshi *et al.*, 2008; Dwork *et al.*, 2012; Kleinberg *et al.*, 2017; Hardt *et al.*, 2016; Zafar *et al.*, 2017; Chouldechova, 2017; Corbett-Davies *et al.*, 2017; Kilbertus *et al.*, 2017; Lipton *et al.*, 2018]. Most of those works focus on classification and propose definitions of fairness based on equating some combinations of the classification outcome (true positives, true negatives, etc.). In this work, we focus on two

classical notions of fairness for the one-shot classification setting: demographic parity (or disparate impact) and equal opportunity (or disparate mistreatment) [Hardt *et al.*, 2016; Zafar *et al.*, 2017]. There are also works on fairness in sequential learning [Joseph *et al.*, 2016; Jabbari *et al.*, 2017; Heidari and Krause, 2018; Valera *et al.*, 2018]. The model in those papers is to sequentially consider each individual and make decision for them, but there is no notion of refining selection through multiple stages by getting additional features.

Closer to our work, a few papers investigate multistage classification/selection without fairness considerations [Senator, 2005; Trapeznikov *et al.*, 2012]. [Schumann *et al.*, 2019] model the interview decisions in hiring as a multi-armed bandit problem and consider getting extra features at a cost for a subset of candidates, but they do not have fairness constraints: they propose an algorithm for their bandit problem and show that it leads to higher diversity than other algorithms.

To the best of our knowledge, our paper is the first that proposes concrete fairness notions for multistage selection and algorithms to maximize utility under fairness constraints. The only other papers discussing fairness in the context of two-stage or composed decision making are [Bower *et al.*, 2017; Dwork and Ilvento, 2019], but they do not model additional features becoming available at the second stage for the subselected individuals, which is the key element of our analysis.

In recent work, [Kleinberg and Raghavan, 2018] consider the problem of selecting a subset of candidates to interview and show that under some condition, imposing diversity may increase utility when there is implicit bias. Their model, however, assumes no statistical knowledge of the features revealed at second stage, and they only maximize the sum of values of subselected candidates (effectively reducing to one-stage). In contrast, we do not consider implicit bias but we do model the second-stage process. Interestingly, our optimal solution also introduces diversity at the first stage selection, but for different reasons.

2 Multistage Selection Framework

2.1 Basic Setting and Notation

Assume that there are n candidates,¹ each described by d features, and consider the following k -stage selection process. At the first stage, we observe some of the features x_1, \dots, x_{d_1} of the n candidates where $d_1 < d$. We then select n_1 of them that “pass” to the second stage. At the second stage, we observe some extra features of these n_1 candidates $x_{d_1+1}, \dots, x_{d_2}$ ($d_1 < d_2$) that were not known at the previous stage. Using the features of both stages, we do a selection, from the n_1 that passed the first stage of $n_2 \leq n_1$ candidates that pass to the next stage, and so on. At the last stage k , we observe all $d_k = d$ features of the n_{k-1} candidates and select n_k among those who passed the stage $k - 1$.

We assume that each candidate is endowed with a *label* $y \in \{0, 1\}$, which encodes whether the candidate is “good” or “bad” according to the purpose of the selection, i.e., if $y = 1$ we would like to have this candidate in our final selection, if

¹We use the term candidates in a generic sense to refer to elements of the initial set that can be selected.

$y = 0$ we would prefer not. The label y is not known until the end and is therefore not available to make the selection.

We assume that the decision maker knows the joint distribution of features and the conditional probability that expresses the probability that the candidate is “good” given all its features. We will denote by $p_{x_1 \dots x_d} = P(x_1, \dots, x_d)$ the probability to observe a specific realization of features and by $p_{x_1 \dots x_d}^{y=1} = P(y = 1 | x_1, \dots, x_d)$ the probability that a candidate is good ($y = 1$) given its features $x_1 \dots x_d$.

2.2 Probabilistic Selection and Budget Constraints

In the following, we will consider a class of selection algorithms that perform a probabilistic selection of candidates. Such an algorithm takes as an input a list of probability values $p_{x_1 \dots x_{d_i}}^{(i|i-1)}$ for all stages $i \in \{1 \dots k\}$ and all possible combination of features. Then, for each candidate that passed stage $i - 1$ and has features $(x_1 \dots x_{d_i})$, the algorithm selects this candidate for the next stage with probability $p_{x_1 \dots x_{d_i}}^{(i|i-1)}$, with the convention that everyone passes stage 0.

For each stage i , we define a binary predictor \hat{y}_i that is equal to 1 if the candidate is selected at stage i (by convention, $\hat{y}_0 = 1$ for all candidates). We assume that, *on average*, the number of candidates that can be selected by the algorithm at stage i is at most $\alpha_i n$ and exactly $\alpha_k n$ for the last stage, with $1 \geq \alpha_1 \geq \dots \geq \alpha_k$. We denote by $\alpha_{-k} = (\alpha_1, \dots, \alpha_{k-1})^T$ the selection sizes of the first $k - 1$ stages.

2.3 Performance Metric

We measure the performance of a given selection algorithm in terms of precision. The precision is the fraction of the selected candidates that indeed were “good” for selection:

$$\text{precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} = P(y = 1 | \hat{y}_k = 1),$$

where the denominator is the number of selected candidates.

The choice of precision may seem arbitrary but it is in fact a very natural metric when the size of the final selection is fixed as in our setting. Indeed, maximizing precision is then equivalent to maximizing most other meaningful metrics as formalized in the next proposition.

Proposition 1. *Assume that the selection size $P(\hat{y}_k = 1)$ is fixed (to α_k). Then maximization of precision is equivalent to maximization of true positive rate, true negative rate, accuracy and f_1 -score; and to minimization of false positive rate and false negative rate.*

Additionally, there are many realistic k -stage selection processes for which precision can be used as a utility metric.

Example 1. *A bank decides to whom it will give entrepreneur loans. The procedure is in two stages: at first, n candidates fill in an application form, and the first d_1 features of each candidate are obtained. Some candidates are then invited for an interview which brings additional features of those candidate that the bank can use for its final decision of selecting n_2 candidates. If the profit of giving a loan to a trustworthy candidate is c_p and if a candidate that does not pay a loan costs c_l , then the average gain can be written as:*

$$U_{\text{bank}} = (c_p + c_l) \cdot n_2 \cdot P(y = 1 | \hat{y}_2 = 1) - n_2 \cdot c_l.$$

In this example, c_p , c_l and selection size n_2 are fixed. Hence, maximizing precision or utility is equivalent.

3 Fairness Notions in Multistage Setting

In this section, we propose new notions of fairness for the multistage selection problem. We assume that there exists, amongst all features that describe candidates, a sensitive feature x_s that indicates whether or not a candidate belongs to a sensitive group that should not be discriminated against.

The literature has introduced multiple definitions of fairness for the single-stage setting (and it is worth mentioning that in most of the cases those fairness criteria cannot be satisfied simultaneously [Chouldechova, 2017]). The most relevant notions in the context of selection problems are *Demographic Parity* (DP) and *Equal Opportunity* (EO). We first recall the definition of these fairness criteria in the traditional setting of single-stage selection. We then extend them to the multistage setting by showing that there are essentially two relevant notions of fairness: *local* and *global* fairness.

3.1 Classical Fairness Notions in Single-Stage

Let \hat{y} be a binary predictor that decides which candidates belong to the selection. The first fairness definition, widely known as demographic parity, states that the predictor \hat{y} is fair if it is statistically independent from x_s .

Definition 1 (Demographic Parity, DP). *The binary predictor \hat{y} satisfies DP with respect to x_s if \hat{y} and x_s are independent:*

$$P(\hat{y} = 1 | x_s = 0) = P(\hat{y} = 1 | x_s = 1). \quad (1)$$

DP does not take into account the actual label y . [Hardt et al., 2016; Zafar et al., 2017] argue that DP is not the most relevant notion of fairness in cases where we have ground truth on the quality of the candidates (which is our case since we assume statistical knowledge of the probabilities of labels). In such cases, one might want to be fair among the candidates that are worth selecting, a metric called Equal Opportunity [Hardt et al., 2016] (an equivalent notion called disparate mistreatment is proposed in [Zafar et al., 2017]):

Definition 2 (Equal Opportunity, EO [Hardt et al., 2016]). *The binary predictor \hat{y} satisfies EO with respect to x_s if \hat{y} and x_s are independent given that $y = 1$:*

$$P(\hat{y} = 1 | y = 1, x_s = 0) = P(\hat{y} = 1 | y = 1, x_s = 1). \quad (2)$$

In the remainder, we systematically consider DP and EO.

3.2 Local and Global Fairness in Multistage

Existing fairness notions apply to single-stage selection, where we have only one binary predictor \hat{y} . In the case of k -stage selection, we have k binary predictors $\hat{y} = (\hat{y}_1, \dots, \hat{y}_k)$. In this section, we develop different notions of fairness that extend existing notions to the k -stage selection setting.

We propose three definitions that we believe correspond to three reasonable notions of fairness. The high-level idea of each definition is depicted on Figure 1. For the sake of brevity of exposition, we present the formal definitions for the demographic parity criterion, the translation to EO (or to any other fairness notion) being straightforward.

The first fairness notion, local fairness 1 (LF1), imposes that the selection be fair at every stage with respect to the set of candidates that reached that stage. In other words the

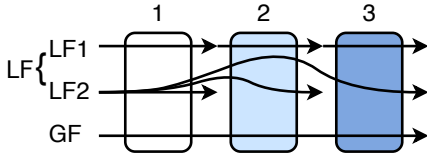


Figure 1: Illustration of the different fairness definitions.

selection of each stage i is fair with respect to the population that “passed” stage $i - 1$.

Definition 3 (Local Fairness 1, LF1). A k -stage selection algorithm satisfies LF1 if (for the case of DP), $\forall i \in \{1, \dots, k\}$:

$$P(\hat{y}_i = 1 | \hat{y}_{i-1} = 1, x_s = 0) = P(\hat{y}_i = 1 | \hat{y}_{i-1} = 1, x_s = 1).$$

The second fairness notion that we propose, local fairness 2 (LF2), prescribes that the selection should be fair at each stage with respect to the initial set of candidates.

Definition 4 (Local Fairness 2, LF2). A k -stage selection algorithm satisfies LF2 if (for the case of DP), $\forall i \in \{1, \dots, k\}$:

$$P(\hat{y}_i = 1 | x_s = 0) = P(\hat{y}_i = 1 | x_s = 1).$$

In the last definition, global fairness (GF), we allow the predictor \hat{y}_i to be unfair at each stage before the last, but we require the final decision \hat{y}_k to be fair with respect to the initial set of candidates.

Definition 5 (Global Fairness, GF). A k -stage selection algorithm satisfies GF if (for the case of DP):

$$P(\hat{y}_k = 1 | x_s = 0) = P(\hat{y}_k = 1 | x_s = 1).$$

Note that the above definitions can be adapted to EO by conditioning on $y = 1$ in all formulas.

3.3 Equivalence between LF1 and LF2

In the following proposition, we show that both notions of local fairness, LF1 and LF2 are equivalent. Therefore in the rest of the paper, we will simply name a multistage selection algorithm that satisfies LF1 (and thus LF2) as a being *locally fair* (LF). An algorithm satisfying the global fairness definition will be called *globally fair* (GF).

Proposition 2 (Relations between fairness notions). *For both DP and EO:*

1. A selection algorithm satisfies LF1 if and only if it satisfies LF2. We call such an algorithm *locally fair* (LF).
2. A locally fair selection algorithm is globally fair (GF).

4 Utility Maximization as a Linear Program

Our goal is to find the binary predictors $(\hat{y}_1, \dots, \hat{y}_k)$ corresponding to stages from 1 to k , respectively, that maximize precision while respecting budget and fairness constraints:

$$\begin{aligned} \max_{\hat{y}_1, \dots, \hat{y}_k} \quad & P(y = 1 | \hat{y}_k = 1) \\ & P(\hat{y}_i = 1) \leq \alpha_i, \quad i \leq k-1 \\ & P(\hat{y}_k = 1) = \alpha_k \\ & f_j(\hat{y}_1, \dots, \hat{y}_k) = 0, \quad j \leq t \end{aligned} \quad (3)$$

where functions $f_j(\cdot)$ of the binary predictors correspond to the fairness constraints we impose. For instance, for a globally fair algorithm (DP) we have only one fairness constraint: $f(\hat{y}_1, \dots, \hat{y}_k) = P(\hat{y}_k = 1 | x_s = 0) - P(\hat{y}_k = 1 | x_s = 1)$.

Using the assumption that the final stage size constraint is $P(\hat{y}_k = 1) = \alpha_k$ we can write the precision as follows:

$$P(y = 1 | \hat{y}_k = 1) = \frac{1}{\alpha_k} \sum_{x_1 \dots x_d} p_{x_1 \dots x_d}^{y=1} p_{x_1 \dots x_d} \prod_{j=1}^k p_{x_1 \dots x_{d_j}}^{(j|j-1)}. \quad (4)$$

Using the notation introduced in Section 2.2, the probability $P(\hat{y}_i = 1)$ that candidate passes stage i is

$$P(\hat{y}_i = 1) = \sum_{x_1 \dots x_d} p_{x_1 \dots x_d} \prod_{j=1}^i p_{x_1 \dots x_{d_j}}^{(j|j-1)}. \quad (5)$$

Hence, the constraints on the selection size $P(\hat{y}_i = 1) \leq \alpha_i$ for $i < k$ and $P(\hat{y}_k = 1) = \alpha_k$ can be expressed using (5).

The fairness constraints can be developed in the same manner, e.g., for the globally fair case (DP):

$$f(\hat{y}_1, \dots, \hat{y}_k) = P(\hat{y}_k = 1 | x_s = 0) - P(\hat{y}_k = 1 | x_s = 1),$$

where $\forall a \in \{0, 1\}$,

$$P(\hat{y}_k = 1 | x_s = a) = \frac{\sum_{x_i, i \neq s} \prod_{j=1}^k p_{x_1 \dots x_{d_j}}^{(j|j-1)} \cdot p_{x_1 \dots x_s = a \dots x_d}}{\sum_{x_i, i \neq s} p_{x_1 \dots x_s = a \dots x_d}}. \quad (6)$$

From (4), we see that the objective is not linear in the variables $p_{x_1 \dots x_{d_j}}^{(j|j-1)}$ due to the product of probabilities. Similarly, we observe from (5) and (6) that the constraints are also not linear in these variables. However, we can show that by using the change of variables $\tilde{p}_{x_1 \dots x_{d_i}}^{(i|i-1)} = \prod_{j=1}^i p_{x_1 \dots x_{d_j}}^{(j|j-1)}$, it can be made linear. This shows that it is possible to compute the variables $p_{x_1 \dots x_{d_j}}^{(j|j-1)}$ that maximize precision (3) using a linear program (LP) (see details in Appendix A), which is key to applicability. It should be noted, however, that the number of variables in (LP) grows exponentially with the number of features.

To distinguish between the different notions of fairness, we will denote by $U_{LF}^*(\alpha_{-k}, \alpha_k)$ and $U_{GF}^*(\alpha_{-k}, \alpha_k)$ the value of the problem (LP)—i.e., the maximum utility—when the fairness constraints correspond to local and global fairness, respectively. Similarly, we will denote by $U_{un}^*(\alpha_{-k}, \alpha_k)$ the optimal precision value when no fairness constraint are imposed (we call it the *unfair* case).

4.1 Solution Properties wrt Budget Constraints

The selection sizes may be related to some budget or to some physical resources of our problem and are crucial parameters. As we show in the next proposition, the optimal utility values are monotonic and concave as functions of budget sizes $\alpha_1, \dots, \alpha_{k-1}$. This property can be useful for budget optimization and is illustrated as well on Figure 2.

Proposition 3 (Monotonicity and concavity). *For $U^* \in \{U_{LF}^*, U_{GF}^*, U_{un}^*\}$ and any fairness constraints that can be expressed as linear homogeneous equations² (such as DP and EO), we have that $U^*(\alpha_{-k}, \alpha_k)$ is*

²See details in Lemma 1 in Appendix A.

1. *non-decreasing and concave with respect to α_{-k}* ;
2. *non-increasing with respect to α_k* .

Note that U^* can be concave or convex or none of the two with respect to α_k , depending on the problem’s parameters.

4.2 The Price of Local Fairness

We are now ready to define our central notion—the *price of local fairness*—that represents the price to pay for being fair at intermediate stages compared to a globally fair solution.

Definition 6 (Price of Local Fairness, $PoLF$). *Let*

$$PoLF(\alpha_{-k}, \alpha_k) = \frac{U_{GF}^*(\alpha_{-k}, \alpha_k)}{U_{LF}^*(\alpha_{-k}, \alpha_k)}.$$

It should be clear that the locally fair algorithm is more constrained than the globally fair. Thus, we have:

$$U_{LF}^*(\alpha_{-k}, \alpha_k) \leq U_{GF}^*(\alpha_{-k}, \alpha_k) \leq U_{un}^*(\alpha_{-k}, \alpha_k).$$

This implies that the values of $PoLF(\alpha_{-k}, \alpha_k)$ are always larger than or equal to 1. Using only the final selection size α_k , it is also possible to compute an upper bound as follows.

Proposition 4 ($PoLF$ bound). *For all (α_{-k}, α_k) , we have:*

$$1 \leq PoLF(\alpha_{-k}, \alpha_k) \leq \min \left(\frac{1}{\alpha_k}, \frac{1}{P(y=1)} \right).$$

For instance, if the final stage selection size is $\alpha_k = 0.3$ (as in our numerical examples), the globally fair algorithm can outperform the locally fair one by a factor at most 3.33. While this bound is probably loose, we will see in our numerical example that the $PoLF$ can be as large as 1.6 on real data.

5 Empirical Analysis

In this section we implement³ the optimization algorithms in order to capture tendencies on real datasets and to provide general insights. We consider the two-stage selection process, since it is the most easily interpretable. Thus, $\alpha_{-k} = \alpha_1$ and $\alpha_k = \alpha_2$. In our experiments we use three datasets: Adult [Dua and Graff, 2017], COMPAS [Larson *et al.*, 2016] and German Credit Data [Dua and Graff, 2017]. We adapt these datasets to our two stage fair selection problem by leaving 6 features, binarizing them (see details in Appendix D) and artificially separating in two stages. We estimate the statistics $p_{x_1 \dots x_d}$ and $p_{x_1 \dots x_d}^{y=1}$ from data. We then use a linear solver for the linear program (LP) that gives us the optimal utility $U^*(\alpha_1, \alpha_2)$ for the fair and unfair cases.

5.1 Analysis of the Price of Local Fairness

We consider three different scenarios: i) the sensitive attribute x_s is observed at the first stage; ii) at the second stage; iii) never used in the selection process. We distinguish these three cases since it could happen that the use of the sensitive attribute x_s in decision making is forbidden at some stages or even at all (by law or other conventions). Our aim is to compare how the price of local fairness behaves in every case.

Let us start with a simple example. We leave 5 features from the Adult dataset: *sex, age, education, relationship*

³All codes are available at https://github.com/vitaly-emelianov/multistage_fairness/

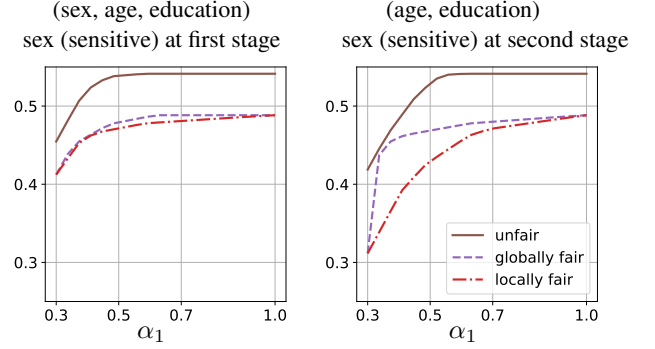


Figure 2: Utility $U^*(\alpha_1, \alpha_2 = 0.3)$ for Adult dataset (DP).

and *native country* and consider the attribute *sex* as sensitive. Figure 2 then shows the values of $U_{\{un, GF, LF\}}^*(\alpha_1, \alpha_2)$ as a function of α_1 for fixed $\alpha_2 = 0.3$ when using the features displayed on top of each subfigure at first stage and the rest at second stage. We make two important observations from this figure. *First*, the value of $PoLF$ can be significant. From Figure 2-(right), we see that for $\alpha_1 \approx 0.33$, the value of $PoLF$ is about 1.3, meaning that the globally fair algorithm achieves 30% larger value of precision than the locally fair. *Second*, the gap between LF and GF algorithms is significantly larger when the sensitive attribute x_s is observed at the second stage.

To show that this behavior is significant we calculate the values of $U_{\{un, GF, LF\}}^*(\alpha_1, \alpha_2)$ for every possible combination $X = \{x_1, \dots, x_5\}$ of 5 features out of 6 as decision variables (x_1, x_2 at first stage and x_3, x_4 at second stage), with one sensitive attribute $x_s = x_5$ that can be observed at the first stage or at the second stage or not observed at all, and for every possible (discretized) value of $\alpha_1 \geq \alpha_2$. Due to space constraints we present our results only for the DP definition of fairness; we emphasize that the observations are robust among the three datasets and the two fairness notions (DP and EO) (see Appendix C for additional results). Figure 3 shows the empirical cumulative distribution functions $\hat{F}_{PoLF}(x)$ of the values of $PoLF$ obtained. We observe that the *price of local fairness is significantly lower when the sensitive attribute x_s is revealed at the first stage* compared to the case where it is revealed later. This is consistent with the observation made on Figure 2. A possible interpretation is that the LF algorithm has to make a conservative decision at the first stage and therefore cannot perform well compared to the GF algorithm that is able to compensate (when the sensitive feature x_s is observed) for the unfair decisions that have been made at the first stage. It is worth mentioning that we have the same observation for a three-stage algorithm: the later we reveal the sensitive attribute, the higher the values of $PoLF$ we obtain (see Appendix C.4).

5.2 Violation of Local Fairness

By definition, a globally fair algorithm can violate fairness constraints at intermediate stages. For a given budget constraints α_1, α_2 , we define the violation of local fairness ($VoLF$) as the absolute value of the fairness constraint violation at the first stage for the optimal globally fair algorithm. For instance, for DP, this quantity equals:

$$VoLF(\alpha_1, \alpha_2) = |P(\hat{y}_1 = 1 | x_s = 0) - P(\hat{y}_1 = 1 | x_s = 1)|.$$

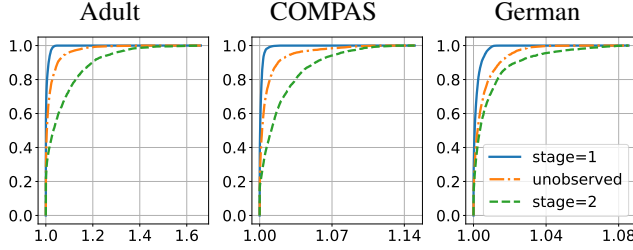


Figure 3: Empirical CDFs of $PoLF$ for all datasets (DP, $\alpha_2 = 0.3$).

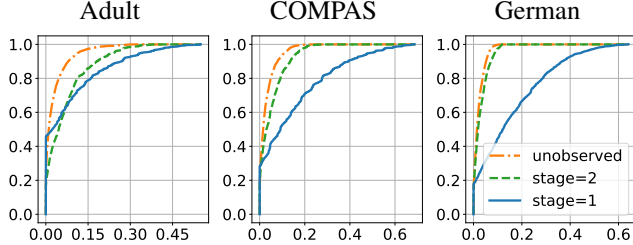


Figure 4: Empirical CDFs of $VoLF$ for all datasets (DP, $\alpha_2 = 0.3$).

Figure 4 shows the empirical cumulative distribution function of violation of fairness $\hat{F}_{VoLF}(x)$ for every value of $\alpha_1 \in [\alpha_2; 1]$ and for every feature combination. We observe that *the later the sensitive feature x_s is revealed (or even not revealed), the more fair at intermediate stages the globally fair algorithm is*. One possible explanation is that an algorithm that cannot observe the sensitive feature x_s at the first stage has to be more “cautious” at every stage to be able to satisfy global fairness since the exact value of sensitive attribute x_s is not available. This observation is again robust among different datasets and notions of fairness.

Finally, on Figure 5 we represent the joint distribution of $PoLF$ and $VoLF$. As mentioned before, the globally fair algorithm is more unfair at the intermediate stages when the sensitive feature x_s is observed from the beginning (left panel), however the price of local fairness we pay in this case is the smallest one. When the sensitive feature x_s is observed at the second stage (middle panel) the globally fair algorithm is more locally fair compared to the previous case, but the value of $PoLF$ is way larger. Finally, when x_s is never observed (right panel) the globally fair algorithm is the “most locally fair” among all three settings. We finally observe that, while most points have either $PoLF$ small (i.e., using a LF algorithm does not lose much) or $VoLF$ small (i.e., the GF algorithm is almost locally fair), there exist some points—when the sensitive feature is observed at the second stage—where both $PoLF$ and $VoLF$ are large; i.e., imposing local fairness even approximately comes at a significant cost.

6 Conclusion

In this work we tackle the problem of multistage selection and the fairness issues it entails. We propose a stylized model based on a probabilistic formulation of the k -stage selection problem with constraints on the number of selected individuals at each stage that should hold in expectation. We introduce two different notions of fairness for the multistage setting: local (under two equivalent variants) and global fairness. Thanks to this framework, we show that maximizing preci-

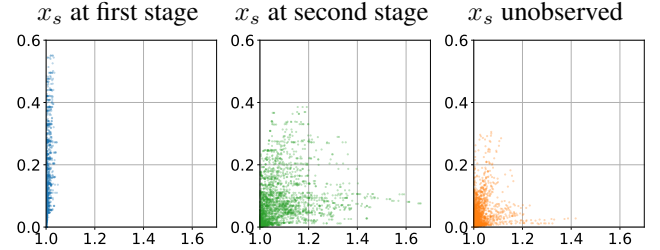


Figure 5: $VoLF$ (y -axis) vs $PoLF$ (x -axis) for **Adult** dataset (DP, $\alpha_2 = 0.3$).

sion under budget and fairness constraints can be done via linear programming, which enables for efficient computation as well as theoretical investigation. In particular, we analyze theoretically and empirically how the utility of locally and globally fair algorithms vary with selection budgets, and we find that globally fair algorithms can lead to non-negligible performance increases compared to locally fair ones.

One of the main findings of our work is that the stage at which the sensitive attribute is revealed greatly affects the difference between the performance of locally and globally fair algorithms: hiding the sensitive feature at early stages tends to make globally fair algorithm more fair at intermediate stages. While locally fair algorithms may be desirable, our results show that local fairness does not come for free. They also show that if a decision maker would like to encourage locally fair selection algorithms, there are essentially two choices: either hide the sensitive feature at the first stage or impose by rules the first stage to be fair.

Our model allows us to provide elegant insights into the fairness questions related to multistage selection, yet it does a number of simplifying assumptions that naturally restrict its direct applicability. *First*, our model ignores the issue that the selection probability at a stage depends on which candidates got selected at the previous stages; i.e., it implicitly makes the approximation that at each stage the number of candidates selected for each feature combination is equal to its expectation. In Appendix E, we show that this approximation becomes exact as n tends to infinity. *Second*, we assume perfect statistical knowledge of the joint distribution of features and label values, without bias. *Third*, we consider only discrete features and use a non-compact representation of the selection probabilities—this allows us to solve the exact selection problem by using an LP formulation. Relaxing these assumptions, in particular using a more compact representation of the selection algorithm (at the cost of a loss of precision) is an interesting direction of future work.

Acknowledgments

This work was supported in part by the French National Research Agency (ANR) through the “Investissements d’avenir” program (ANR-15-IDEX-02) and through grant ANR-16-TERC0012; by the Alexander von Humboldt Foundation; and by a European Research Council (ERC) Advanced Grant for the project “Foundations for Fair Social Computing” funded under the European Union’s Horizon 2020 Framework Programme (grant agreement no. 789373). The authors also thank Roland Hildebrand for helpful technical suggestions.

References

- [Bower *et al.*, 2017] Amanda Bower, Sarah N. Kitchen, Laura Niss, Martin J. Strauss, Alex Vargo, and Suresh Venkatasubramanian. Fair pipelines. In *Proceedings of the 4th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT-ML)*, 2017.
- [Chouldechova, 2017] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.
- [Corbett-Davies *et al.*, 2017] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 797–806, 2017.
- [Dua and Graff, 2017] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.
- [Dwork and Ilvento, 2019] Cynthia Dwork and Christina Ilvento. Fairness under composition. In *Proceedings of the 10th conference on Innovations in Theoretical Computer Science (ITCS)*, pages 33:1–33:20, 2019.
- [Dwork *et al.*, 2012] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd conference on Innovations in Theoretical Computer Science Conference (ITCS)*, pages 214–226, 2012.
- [Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)*, pages 3323–3331, 2016.
- [Heidari and Krause, 2018] Hoda Heidari and Andreas Krause. Preventing disparate treatment in sequential decision making. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2248–2254, 2018.
- [Jabbari *et al.*, 2017] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1617–1626, 2017.
- [Joseph *et al.*, 2016] Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)*, pages 325–333, 2016.
- [Kilbertus *et al.*, 2017] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, pages 656–666, 2017.
- [Kleinberg and Raghavan, 2018] Jon Kleinberg and Manish Raghavan. Selection problems in the presence of implicit bias. In *Proceedings of the 9th conference on Innovations in Theoretical Computer Science (ITCS)*, pages 33:1–33:17, 2018.
- [Kleinberg *et al.*, 2017] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th conference on Innovations in Theoretical Computer Science (ITCS)*, pages 43:1–43:23, 2017.
- [Lambrecht and Tucker, 2018] Anja Lambrecht and E. Tucker, Catherine. Algorithmic bias? an empirical study into apparent gender-based discrimination in the display of stem career ads, March 2018. Available at SSRN: <https://ssrn.com/abstract=2852260>.
- [Larson *et al.*, 2016] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How We Analyzed the COMPAS Recidivism Algorithm, 2016. ProPublica, <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.
- [Lipton *et al.*, 2018] Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating ml’s impact disparity require treatment disparity? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS)*, pages 8125–8135, 2018.
- [Pedreshi *et al.*, 2008] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 560–568, 2008.
- [Perry *et al.*, 2013] Walter L. Perry, Brian McInnis, Carter C. Price, Susan C. Smith, and John S. Hollywood. *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*. Rand Corporation, 2013.
- [Rohatgi and Saleh, 2015] Vijay K. Rohatgi and A.K.M.E. Saleh. *An Introduction to Probability and Statistics*. Wiley Series in Probability and Statistics. Wiley, 2015.
- [Schumann *et al.*, 2019] Candice Schumann, Samsara N. Counts, Jeffrey S. Foster, and John P. Dickerson. The diverse cohort selection problem. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 601–609, 2019.
- [Senator, 2005] Ted E. Senator. Multi-stage classification. In *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM)*, pages 386–393, 2005.
- [Trapeznikov *et al.*, 2012] Kirill Trapeznikov, Venkatesh Saligrama, and David Castañón. Multi-stage classifier design. In *Proceedings of the Asian Conference on Machine Learning*, pages 459–474, 2012.
- [Valera *et al.*, 2018] Isabel Valera, Adish Singla, and Manuel Gomez Rodriguez. Enhancing the accuracy and fairness of human decision making. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS)*, pages 1774–1783, 2018.
- [Zafar *et al.*, 2017] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web (WWW)*, pages 1171–1180, 2017.

Appendix A Utility Maximization via Linear Programming

In this appendix, we formally justify that maximizing precision under budget constraints and fairness constraints can be done via linear programming. The following proposition shows how to do that with only budget constraints:

Proposition 5 (Utility maximization as a linear program). *Let, for all $i \in \{1, \dots, k\}$ and all $x_1 \dots x_{d_i}$,*

$$p_{x_1 \dots x_{d_i}}^{(i|i-1)} = \begin{cases} \tilde{p}_{x_1 \dots x_{d_i}}^{(i|i-1)} / \tilde{p}_{x_1 \dots x_{d_{i-1}}}^{(i-1|i-2)}, & \text{if } \tilde{p}_{x_1 \dots x_{d_{i-1}}}^{(i-1|i-2)} \neq 0 \\ 0, & \text{otherwise,} \end{cases}$$

where the variables $\tilde{p}_{x_1 \dots x_{d_i}}^{(i|i-1)}$ are solutions of the linear program

$$\begin{aligned} \max_{\tilde{p}_{x_1 \dots x_{d_i}}^{(i|i-1)}} \quad & \frac{1}{\alpha_k} \sum_{x_1 \dots x_d} p_{x_1 \dots x_d}^{y=1} \cdot p_{x_1 \dots x_d} \cdot \tilde{p}_{x_1 \dots x_{d_k}}^{(k|k-1)} \\ \text{s.t.} \quad & \sum_{x_1 \dots x_d} p_{x_1 \dots x_d} \cdot \tilde{p}_{x_1 \dots x_{d_i}}^{(i|i-1)} \leq \alpha_i, \quad i < k, \\ & \sum_{x_1 \dots x_d} p_{x_1 \dots x_d} \cdot \tilde{p}_{x_1 \dots x_{d_k}}^{(k|k-1)} = \alpha_k, \\ & 0 \leq \tilde{p}_{x_1 \dots x_{d_1}}^{(1|0)} \leq 1, \\ & 0 \leq \tilde{p}_{x_1 \dots x_{d_i}}^{(i|i-1)} \leq \tilde{p}_{x_1 \dots x_{d_{i-1}}}^{(i-1|i-2)}, \quad 1 < i \leq k. \end{aligned} \quad (7)$$

Then the $p_{x_1 \dots x_{d_i}}^{(i|i-1)}$ are solutions of (3) without any fairness constraint.

Proof. Using (4)–(5), we can rewrite problem (3) as

$$\begin{aligned} \max_{p_{x_1 \dots x_{d_i}}^{(i|i-1)}} \quad & \frac{1}{\alpha_k} \sum_{x_1 \dots x_d} p_{x_1 \dots x_d}^{y=1} \cdot p_{x_1 \dots x_d} \cdot \prod_{i=1}^k p_{x_1 \dots x_{d_i}}^{(i|i-1)} \\ \text{s.t.} \quad & \sum_{x_1 \dots x_d} p_{x_1 \dots x_d} \cdot \prod_{j=1}^i p_{x_1 \dots x_{d_j}}^{(j|j-1)} \leq \alpha_i, \quad i < k, \\ & \sum_{x_1 \dots x_d} p_{x_1 \dots x_d} \cdot \prod_{j=1}^k p_{x_1 \dots x_{d_j}}^{(j|j-1)} = \alpha_k, \\ & 0 \leq p_{x_1 \dots x_{d_i}}^{(i|i-1)} \leq 1, \quad 1 \leq i \leq k. \end{aligned} \quad (8)$$

Let us define the new variables

$$\tilde{p}_{x_1 \dots x_{d_i}}^{(i|i-1)} = \prod_{j=1}^i p_{x_1 \dots x_{d_j}}^{(j|j-1)}. \quad (9)$$

By substitution, we get the linear program (7). Hence, assuming that $\tilde{p}_{x_1 \dots x_{d_i}}^{(i|i-1)}$ are solutions of (7), any $p_{x_1 \dots x_{d_i}}^{(i|i-1)}$ such that (9) is satisfied (as is the case for the $p_{x_1 \dots x_{d_i}}^{(i|i-1)}$ defined in the proposition) is a solution of (3). \square

In the following lemma, we then show that fairness constraints can also be written as linear homogeneous equations in terms of the transformed variables $\mathbf{p} = (\tilde{p}_{x_1 \dots x_{d_i}}^{(i|i-1)})^T$.

Lemma 1 (Linearity of fairness constraints). *For both local and global fairness, and for both EO and DP, there exists a matrix \mathbf{F} such that the fairness constraint can be expressed as*

$$\mathbf{F}\mathbf{p} = \mathbf{0}.$$

Proof. We present the proof for demographic parity, the idea is the same for equal opportunity. Let us consider the fairness constraint corresponding to stage i , $1 \leq i \leq k$:

$$P(\hat{y}_i = 1 | x_s = 0) = P(\hat{y}_i = 1 | x_s = 1).$$

By expanding the left side, we obtain:

$$P(\hat{y}_i = 1 | x_s = 0) = \frac{\sum_{x_i, i \neq s} \tilde{p}_{x_1 \dots x_s=0 \dots x_{d_i}}^{(i|i-1)} \cdot p_{x_1 \dots x_s=0 \dots x_d}}{\sum_{x_i, i \neq s} p_{x_1 \dots x_s=0 \dots x_d}}.$$

Recall that $p_{x_1 \dots x_s=0 \dots x_d}$ is a fixed parameter and not a decision variable. Thus, for both local and global fairness, the fairness constraint (equality of the probabilities for $x_s = 0$ and $x_s = 1$) can be represented in the form $\mathbf{F}\mathbf{p} = \mathbf{0}$ for an appropriate \mathbf{F} simply by moving all terms on the left side of the equality. \square

Appendix B Missing Proofs

In this appendix, we provide the proofs of all results stated in the paper. We start by introducing notation that will be used throughout the proofs.

Notation

To ease the exposition, we introduce the following matrix notation for the problem (7).

- *Selection probabilities \mathbf{p} .* We concatenate all the selection probabilities in a single vector:

$$\mathbf{p} = (\tilde{p}_{x_1 \dots x_{d_i}}^{(i|i-1)})^T,$$

where $\tilde{p}_{x_1 \dots x_{d_i}}^{(i|i-1)}$ is the vector of selection probabilities at stage i for all possible values of $x_1 \dots x_{d_i}$ (whose size depends on i).

- *Constraints set $C_{\alpha_{-k}, \alpha_k}$.* We have two types of constraints in problem (7).

1. The constraints that correspond to selection sizes α_i , $i = 1, \dots, k$. We separate them such that $\mathbf{A}\mathbf{p} \leq \alpha_{-k}$ corresponds to selection at first $k-1$ stages, so

$$\alpha_{-k} = (\alpha_1, \dots, \alpha_{k-1})^T.$$

The constraint $\mathbf{b}^T \mathbf{p} = \alpha_k$ corresponds to selection at the last stage, where we require a strict equality.

2. The constraints that do not depend on selection sizes are written in a form of $\mathbf{D}\mathbf{p} \leq \delta$ for an appropriate \mathbf{D} , where $\delta = (1, \dots, 1, 0, \dots, 0)^T$: 1's in δ correspond to constraints $0 \leq \tilde{p}_{x_1 \dots x_{d_1}}^{(1|0)} \leq 1$ and 0's correspond to constraints $0 \leq \tilde{p}_{x_1 \dots x_{d_i}}^{(i|i-1)} \leq \tilde{p}_{x_1 \dots x_{d_{i-1}}}^{(i-1|i-2)}$.

Thus, we write every constraint in matrix form and introduce the following compactly formed constraint set:

$$C_{\alpha_{-k}, \alpha_k} = \{\mathbf{p} \in [0, 1]^d : \mathbf{A}\mathbf{p} \leq \alpha_{-k}, \mathbf{b}^T \mathbf{p} = \alpha_k, \mathbf{D}\mathbf{p} \leq \delta\}.$$

- *Utility function $U_{\alpha_{-k}, \alpha_k}(\mathbf{p})$,* can be written as

$$U_{\alpha_{-k}, \alpha_k}(\mathbf{p}) = \frac{1}{\alpha_k} \mathbf{c}^T \mathbf{p},$$

where $\mathbf{c} = (p_{x_1 \dots x_d}^{y=1} \cdot p_{x_1 \dots x_d})^T$.

Proposition 5 and Lemma 1 show that the problem of maximizing precision (or equivalently, other metrics, see Proposition 1) can be solved through a linear program when the selection sizes sizes

(α_{-k}, α_k) are given constants. This shows that the utility maximization problem in general form can be written as:

$$U^*(\alpha_{-k}, \alpha_k) = \max_{\mathbf{p} \in C_{\alpha_{-k}, \alpha_k} \cap C_f} \frac{1}{\alpha_k} \mathbf{c}^T \mathbf{p}, \quad (10)$$

where

$$C_{\alpha_{-k}, \alpha_k} = \{\mathbf{p} \in [0, 1]^d : \mathbf{A}\mathbf{p} \leq \alpha_{-k}, \mathbf{b}^T \mathbf{p} = \alpha_k, \mathbf{D}\mathbf{p} \leq \delta\},$$

$$C_f = \{\mathbf{p} \in [0, 1]^d : \mathbf{F}\mathbf{p} = \mathbf{0}\}.$$

Proof of Proposition 1

We prove the equivalence only for accuracy (denoted ACC); the proof for other metrics follows the same idea. By expanding ACC , we obtain:

$$\begin{aligned} ACC &= P(\hat{y}_k = y) = P(\hat{y}_k = 1, y = 1) + P(\hat{y}_k = 0, y = 0) \\ &= P(\hat{y}_k = 1, y = 1) + (P(y = 0) - P(\hat{y}_k = 1, y = 0)) \\ &= 2 \cdot P(\hat{y}_k = 1, y = 1) + P(y = 0) - P(\hat{y}_k = 1) \\ &= 2 \cdot P(\hat{y}_k = 1) \cdot P(y = 1 | \hat{y}_k = 1) \\ &\quad + P(y = 0) - P(\hat{y}_k = 1). \end{aligned}$$

Since the terms $P(\hat{y}_k = 1)$ and $P(y = 0)$ are constant, maximization of precision is equivalent to maximization of ACC .

Proof of Proposition 2

(1) We present the proof only for demographic parity, the proof for equal opportunity follows the same idea. We do the proof by induction. First consider a 2-stage selection algorithm $\hat{y} = (\hat{y}_1, \hat{y}_2)$. By considering the following quantity:

$$\begin{aligned} P(\hat{y}_2 = 1 | \hat{y}_1 = 1, x_s = 0) &= \frac{P(\hat{y}_2 = 1, \hat{y}_1 = 1, x_s = 0)}{P(\hat{y}_1 = 1, x_s = 0)} \\ &= \frac{P(\hat{y}_2 = 1, \hat{y}_1 = 1, x_s = 0) + \overbrace{P(\hat{y}_2 = 1, \hat{y}_1 = 0, x_s = 0)}^{=0}}{P(\hat{y}_1 = 1, x_s = 0)} \\ &= \frac{P(\hat{y}_2 = 1, x_s = 0)}{P(\hat{y}_1 = 1 | x_s = 0) P(x_s = 0)} = \frac{P(\hat{y}_2 = 1 | x_s = 0)}{P(\hat{y}_1 = 1 | x_s = 0)}, \end{aligned}$$

the fairness constraint for LF1 at the second stage is:

$$\frac{P(\hat{y}_2 = 1 | x_s = 0)}{P(\hat{y}_1 = 1 | x_s = 0)} = \frac{P(\hat{y}_2 = 1 | x_s = 1)}{P(\hat{y}_1 = 1 | x_s = 1)}.$$

Since we impose fairness at the first stage, then $P(\hat{y}_1 = 1 | x_s = 0) = P(\hat{y}_1 = 1 | x_s = 1)$, so the condition above is equivalent to

$$P(\hat{y}_2 = 1 | x_s = 0) = P(\hat{y}_2 = 1 | x_s = 1),$$

that is exactly the second constraint for the LF2 notion. Thus, the statement is true for a 2 stage selection algorithm.

Second, assuming that the statement is true for $\hat{y} = (\hat{y}_1, \dots, \hat{y}_i)$, $i > 2$, let us consider the $(i + 1)$ -stage selection algorithm. By analogy, considering the quantity

$$\begin{aligned} P(\hat{y}_{i+1} = 1 | \hat{y}_i = 1, x_s = 0) &= \frac{P(\hat{y}_{i+1} = 1, \hat{y}_i = 1, x_s = 0)}{P(\hat{y}_i = 1, x_s = 0)} \\ &= \frac{P(\hat{y}_{i+1} = 1, x_s = 0)}{P(\hat{y}_i = 1 | x_s = 0) P(x_s = 0)} = \frac{P(\hat{y}_{i+1} = 1 | x_s = 0)}{P(\hat{y}_i = 1 | x_s = 0)} \end{aligned}$$

we obtain that the fairness constraint for LF1 at the stage $i + 1$ is

$$\frac{P(\hat{y}_{i+1} = 1 | x_s = 0)}{P(\hat{y}_i = 1 | x_s = 0)} = \frac{P(\hat{y}_{i+1} = 1 | x_s = 1)}{P(\hat{y}_i = 1 | x_s = 1)}.$$

As $P(\hat{y}_i = 1 | x_s = 0) = P(\hat{y}_i = 1 | x_s = 1)$ by the assumption of induction, we have $P(\hat{y}_{i+1} = 1 | x_s = 0) = P(\hat{y}_{i+1} = 1 | x_s = 1)$.

The point (2) follows from the definitions of LF2 and GF, since the problem GF is less constrained than LF2.

Proof of Proposition 3

(1) For a given α_{-k} , assume that the program attains its maximum $U^*(\alpha_{-k}, \alpha_k)$ at point \mathbf{p} and let $\alpha'_{-k} \geq \alpha_{-k}$, where \geq is meant component-wise.

By setting $\mathbf{p}' = \mathbf{p}$, we obtain that $\mathbf{p}' \in C_{\alpha'_{-k}, \alpha_k}$ and thus:

$$U^*(\alpha_{-k}, \alpha_k) = U_{\alpha_{-k}, \alpha_k}(\mathbf{p}) = U_{\alpha'_{-k}, \alpha_k}(\mathbf{p}') \leq U^*(\alpha'_{-k}, \alpha_k).$$

Let us consider a problem, when $\alpha_{-k} = \alpha'_{-k}$, it attains its maximum $U^*(\alpha'_{-k}, \alpha_k)$ at the point \mathbf{p}' . Analogously, let for the second problem $\alpha_{-k} = \alpha''_{-k}$, and it attains its maximum $U^*(\alpha''_{-k}, \alpha_k)$ at the point \mathbf{p}'' . Then for any $\lambda \in [0, 1]$ the point $\lambda \mathbf{p}' + (1 - \lambda) \mathbf{p}'' \in C_{\lambda \alpha'_{-k} + (1 - \lambda) \alpha''_{-k}, \alpha_k}$ and:

$$\begin{aligned} \lambda U^*(\alpha', \alpha_k) + (1 - \lambda) U^*(\alpha'', \alpha_k) &= \lambda \frac{1}{\alpha_k} \mathbf{c}^T \mathbf{p}' \\ &\quad + (1 - \lambda) \frac{1}{\alpha_k} \mathbf{c}^T \mathbf{p}'' \\ &= \frac{1}{\alpha_k} \mathbf{c}^T (\lambda \mathbf{p}' + (1 - \lambda) \mathbf{p}'') \\ &= U_{\lambda \alpha'_{-k} + (1 - \lambda) \alpha''_{-k}, \alpha_k}(\lambda \mathbf{p}' + (1 - \lambda) \mathbf{p}'') \\ &\leq U^*(\lambda \alpha'_{-k} + (1 - \lambda) \alpha''_{-k}, \alpha_k). \end{aligned}$$

(2) For a given α_k , assume that the program attains its maximum $U^*(\alpha_{-k}, \alpha_k)$ at point \mathbf{p} . Let $\alpha'_k = \alpha_k / \gamma$, where $\gamma \in [1, +\infty)$. Then consider $\mathbf{p}' = \mathbf{p} / \gamma$. We have $\mathbf{p}' \in C_{\alpha_{-k}, \alpha'_k}$ and $\mathbf{p}' \in C_f$ and:

$$\begin{aligned} U^*(\alpha_{-k}, \alpha_k) &= U_{\alpha_{-k}, \alpha_k}(\mathbf{p}) = \frac{1}{\alpha_k} \mathbf{c}^T \mathbf{p} \\ &= \frac{1}{\alpha_k / \gamma} \mathbf{c}^T \mathbf{p} / \gamma = U_{\alpha_{-k}, \alpha'_k}(\mathbf{p}') \leq U^*(\alpha_{-k}, \alpha'_k). \end{aligned}$$

Proof of Proposition 4

Let us consider the trivial locally fair algorithm. It selects candidates randomly with probability α_1 at the first stage and with probability $\alpha_i / \alpha_{i-1}, \forall 1 < i \leq k$. The utility of such random algorithm is equal to $U_{random}(\alpha_{-k}, \alpha_k) = P(y = 1)$. It is obvious by definition that

$$U_{random}(\alpha_{-k}, \alpha_k) \leq U_{LF}^*(\alpha_{-k}, \alpha_k) \leq U_{GF}^*(\alpha_{-k}, \alpha_k).$$

To obtain an upper bound of $U_{GF}^*(\alpha_{-k}, \alpha_k)$ we suppose that all features are available for the selection, meaning that $\alpha_i = 1, \forall i < k$. Then $U_{GF}^*(\alpha_{-k}, \alpha_k) \leq U_{un}^*(\alpha_1 = 1, \dots, \alpha_{k-1} = 1, \alpha_k) \leq \min(P(y = 1) / \alpha_k, 1)$. Thus,

$$\begin{aligned} PoLF(\alpha_{-k}, \alpha_k) &\leq \frac{\min(P(y = 1) / \alpha_k, 1)}{P(y = 1)} \\ &= \min\left(\frac{1}{P(y = 1)}, \frac{1}{\alpha_k}\right). \end{aligned}$$

Appendix C Additional experimental results

In this appendix, we provide additional experimental results that support the claims in the paper—in particular by reproducing the curves in the paper for other datasets and other fairness metrics.

C.1 The Price of Local Fairness ($PoLF$)

Figure 6 displays the CDF of $PoLF$ as in Figure 3 but including the results for EO as well.

C.2 The Violation of Local Fairness ($VoLF$)

Figure 7 displays the CDF of $VoLF$ as in Figure 7 but including the results for EO as well.

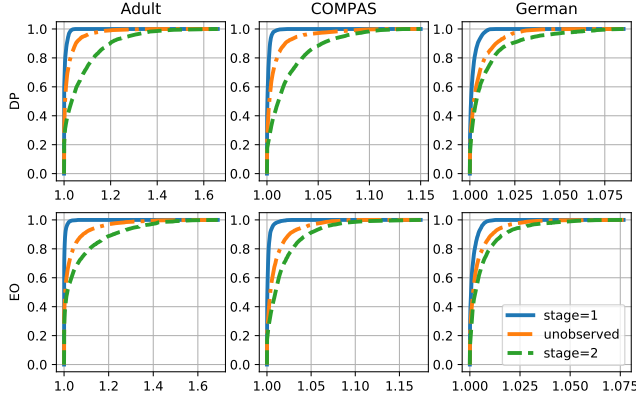


Figure 6: The cumulative distribution function of $PoLF$ for $\alpha_2 = 0.3$.

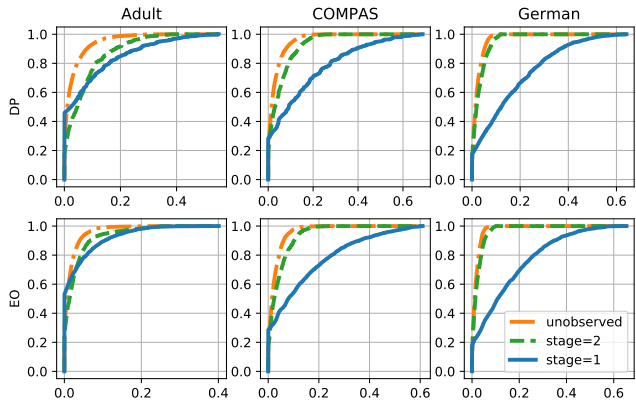


Figure 7: The cumulative distribution function of $VoLF$ for $\alpha_2 = 0.3$.

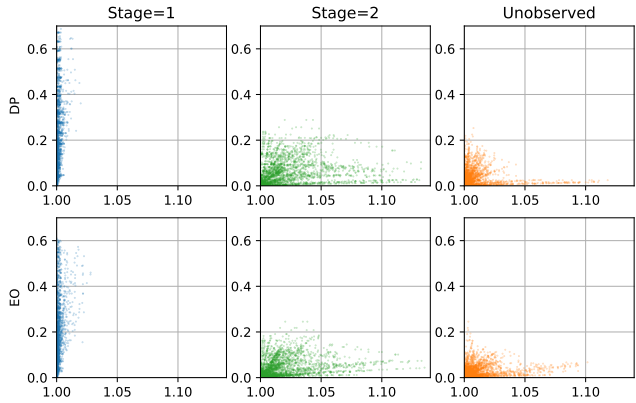


Figure 8: Joint distribution of $VoLF$ and $PoLF$ for COMPAS dataset and $\alpha_2 = 0.3$.

C.3 $VoLF$ vs $PoLF$ for Various Datasets

Figures 8 and 9 display the joint distribution of $VoLF$ (y -axis) and $PoLF$ (x -axis) as in Figure 5 but for the other two datasets: COMPAS and German respectively.

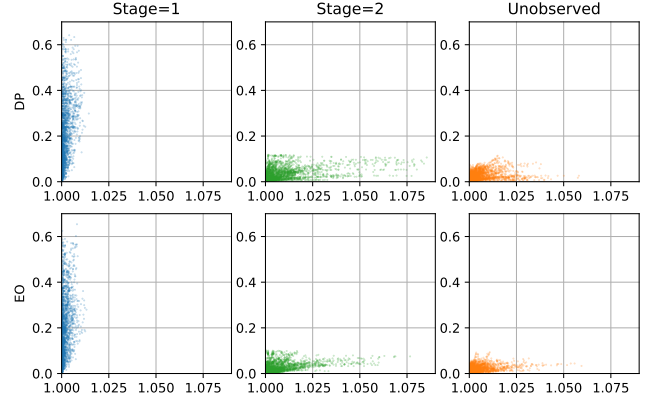


Figure 9: Joint distribution of $VoLF$ and $PoLF$ for German dataset and $\alpha_2 = 0.3$.

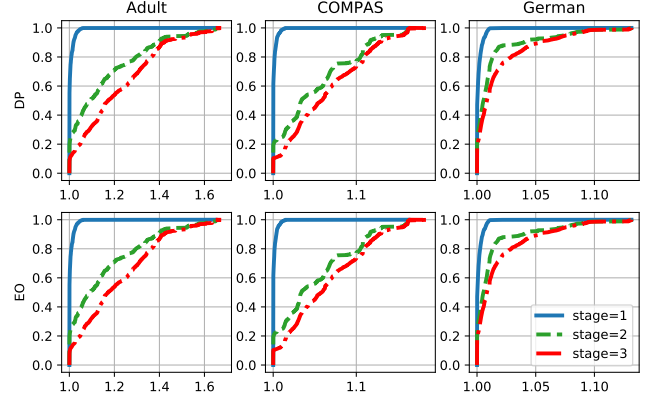


Figure 10: The cumulative distribution function of $PoLF$ of 3-stage algorithm for $\alpha_3 = 0.3$.

C.4 $PoLF$ of 3-stage algorithm

In this subsection we present the results for $PoLF$ of 3-stage selection algorithm. The procedure to calculate the $PoLF$ values is similar to the one we used for two-stage algorithm. We suppose that we observe only one feature at every stage, we suppose that the one of the rest features is a sensitive x_s and consider the cases when it is observed at first, second or third stage of selection process. We calculate the value of $PoLF$ for every possible 4 feature combinations (three decision variables and one being sensitive) out of 6 and for every discretized value of α_1 and α_2 , such that $\alpha_3 = 0.3 \leq \alpha_2 \leq \alpha_1$. Figure 10 displays the empirical CDFs of $PoLF$ for 3-stage algorithm. The observations are the same as in two-stage case: later the sensitive attribute x_s is revealed, larger is the price of imposing local constraints.

On Figure 11 we show the joint distribution of $PoLF$ when the x_s (shown on top of each subfigure) is observed at the first stage (we call it $PoLF_1$) and $PoLF$ when the x_s is observed at the second stage ($PoLF_2$) for Adult dataset. We observe that the value of $PoLF_1$ is sufficiently smaller than the corresponding value of $PoLF_2$.

On Figure 12 we show the joint distribution of $PoLF$ when the x_s is observed at the second stage (we call it $PoLF_2$) and $PoLF$ when the x_s is observed at the third stage ($PoLF_3$) for Adult dataset. We again observe that the value of $PoLF_2$ is smaller than the corresponding value of $PoLF_3$, since the most of the points lie

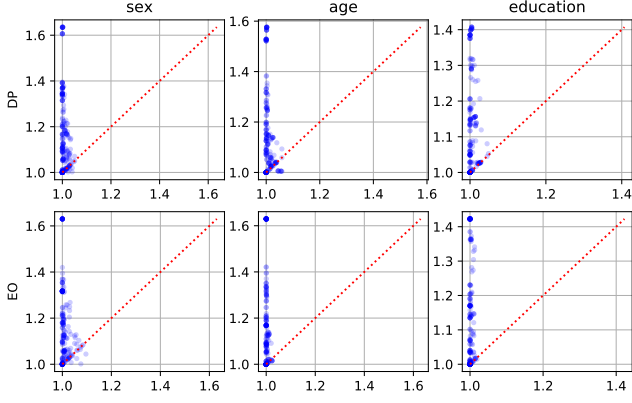


Figure 11: Joint distribution of $PoLF_1$ and $PoLF_2$ of 3-stage algorithm for Adult dataset and $\alpha_3 = 0.3$.

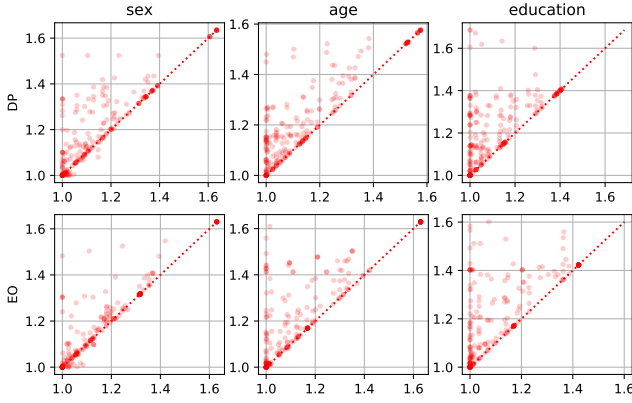


Figure 12: Joint distribution of $PoLF_2$ and $PoLF_3$ of 3-stage algorithm for Adult dataset and $\alpha_3 = 0.3$.

above the diagonal line which is marked as dashed red line.

On Figures 13-16 we display the joint distributions in the same manner as on Figures 11 and 12 but for COMPAS and German Credit datasets.

Appendix D Data Description

As mentioned in the paper, we binarize all features of datasets. In this appendix we describe this step in more detail.

D.1 Adult dataset

In the Adult dataset from the UCI repository [Dua and Graff, 2017] there are 48842 candidates, each described by 14 features. The label *income* denotes if candidate gains more than 50.000 dollars annually. For all our experiments we binarize and leave only the 6 following features: *sex* (is male), *age* (is above 35), *native-country* (from the EU or US), *education* (has Bachelor or Master degree), *hours-per-week* (works more than 35 hours per week) and *relationship* (is married).

D.2 COMPAS Dataset

The COMPAS dataset [Larson *et al.*, 2016] is a dataset that is used to train the COMPAS algorithm. It contains information about prisoners, such as their name, gender, age, race, start of the sentence, end

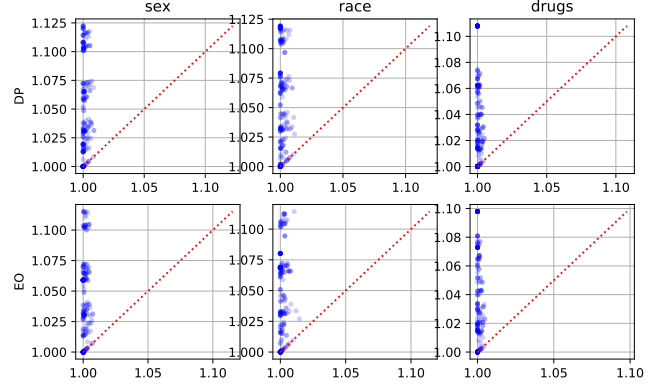


Figure 13: Joint distribution of $PoLF_1$ and $PoLF_2$ of 3-stage algorithm for COMPAS dataset and $\alpha_3 = 0.3$.

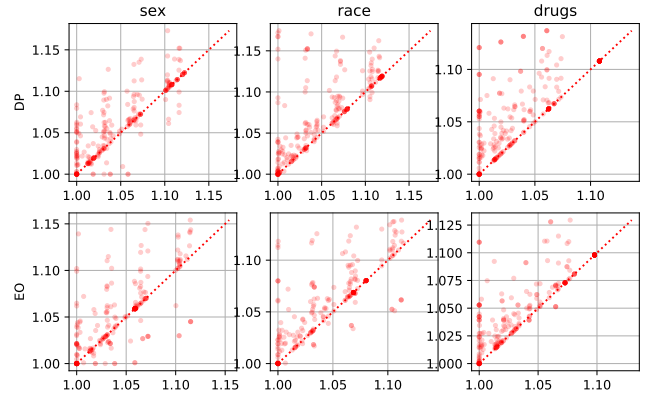


Figure 14: Joint distribution of $PoLF_2$ and $PoLF_3$ of 3-stage algorithm for COMPAS dataset and $\alpha_3 = 0.3$.

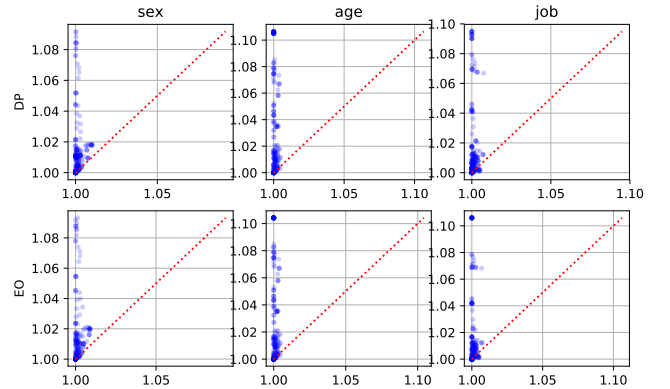


Figure 15: Joint distribution of $PoLF_1$ and $PoLF_2$ of 3-stage algorithm for German Credit dataset and $\alpha_3 = 0.3$.

of the sentence, charge description etc. and a label $y=\text{recidivism}$, that is $y = 1$ if person is likely to reoffend and 0, otherwise.

We prepare original COMPAS dataset for our means by selecting statistics only for Caucasian and African-American defendants, leaving only 6 features and binarizing them. The features that we use are following: *sex* (is male), *young* (younger than 25), *old* (older

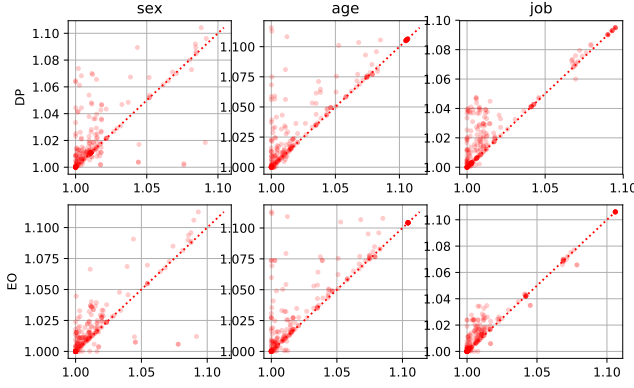


Figure 16: Joint distribution of $PoLF_2$ and $PoLF_3$ of 3-stage algorithm for German Credit dataset and $\alpha_3 = 0.3$.

than 45), *long sentence* (sentence was longer than 30 days), *drugs* (the arrest was due to selling or possessing drugs), *race* (is Caucasian).

D.3 German Dataset

The German Credit data from [Dua and Graff, 2017] contains information about applicants for credit. As with other datasets, we binarize feature values. The label feature $y=returns$ shows if applicant paid for his loan, and we binarize and use 6 features: *job* (is employed), *housing* (owns house), *sex* (is male) *savings* (greater than 500 DM), *credit history* (all credits paid back duly), *age* (older than 50).

Appendix E Utility Maximization In Limit Of $n \rightarrow \infty$

In this appendix we provide an intuition on multistage selection process in limit of infinitely large n . In short, the optimal candidate selection problem with finite number of candidates n appears to be a k -stage stochastic optimization problem which is difficult to solve exactly. By letting the number of candidates n to be infinitely large allows us to reformulate the problem in a much simpler manner such that we are able to find an optimal selection probabilities easily.

To prove the statements in this appendix we will exploit the two following classical results from the probability theory, see [Rohatgi and Saleh, 2015].

Lemma 2 (Chebyshev-Bienaymé inequality). *Let X be a random variable, then*

$$P(|X - EX| \geq \varepsilon) \leq \frac{Var(X)}{\varepsilon^2}.$$

Lemma 3 (Properties of convergence in probability). *Let X_n and Y_n be sequences of random variables.*

1. If $X_n \xrightarrow{P} X$ and a is a constant, then $aX_n \xrightarrow{P} aX$.
2. If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n + Y_n \xrightarrow{P} X + Y$.
3. If $X_n \xrightarrow{P} X$, then $1/X_n \xrightarrow{P} 1/X$.
4. If $X_n \xrightarrow{L} X$ and $|X_n - Y_n| \xrightarrow{P} 0$, then $Y_n \xrightarrow{L} X$.

The following lemma gives us the limit on the proportion of selected at stage i candidates as $n \rightarrow \infty$.

Lemma 4. *Let by $n_{x_1 \dots x_{d_i}}^{(i)}$ denote the number of candidates having features $x_1 \dots x_{d_i}$ that are selected at the stage i , then for $n \rightarrow \infty$:*

$$\frac{n_{x_1 \dots x_{d_i}}^{(i)}}{n} \xrightarrow{L} p_{x_1 \dots x_{d_i}} \prod_{j=1}^i p_{x_1 \dots x_{d_j}}^{(j|j-1)}.$$

Before proving the above lemma let us define the budget $B_n(i)$ at the stage i as $B_n(i) = \frac{1}{n} \sum_{x_1 \dots x_{d_i}} n_{x_1 \dots x_{d_i}}^{(i)}$. Then using property 2 from Lemma 3 and Lemma 4 we obtain that $B_n(i) \xrightarrow{L} \sum_{x_1 \dots x_{d_i}} p_{x_1 \dots x_{d_i}} \prod_{j=1}^i p_{x_1 \dots x_{d_j}}^{(j|j-1)}$.

The precision is the proportion of good candidates among selected then using Lemma 3, the argument in Lemma 4 and fact the the final stage selection size is fixed to α_k , the precision converges in law to $\frac{1}{\alpha_k} \sum_{x_1 \dots x_d} p_{x_1 \dots x_d} p_{x_1 \dots x_d}^{y=1} \prod_{j=1}^k p_{x_1 \dots x_{d_j}}^{(j|j-1)}$ as n goes to infinity. Hence, the equations (4)–(5) hold as the number of candidates $n \rightarrow \infty$. Let us prove Lemma 4 by induction on stage number i .

Proof. Base of induction. Before we do any selection:

$$n_{x_1 \dots x_{d_1}}^{(0)} \sim \text{Bin}\left(n, p_{x_1 \dots x_{d_1}}\right),$$

then using Chebyshev–Bienaymé inequality:

$$P\left(\left|\frac{n_{x_1 \dots x_{d_1}}^{(0)}}{n} - \frac{n \cdot p_{x_1 \dots x_{d_1}}}{n}\right| \geq \varepsilon\right) \leq \frac{np_{x_1 \dots x_{d_1}}(1 - p_{x_1 \dots x_{d_1}})}{\varepsilon^2 n^2} \leq \frac{1}{\varepsilon^2 n}$$

hence $\frac{n_{x_1 \dots x_{d_1}}^{(0)}}{n} \xrightarrow{P} p_{x_1 \dots x_{d_1}}$, $n \rightarrow \infty$.

After, when we perform the selection at the first stage:

$$n_{x_1 \dots x_{d_1}}^{(1)} | n_{x_1 \dots x_{d_1}}^{(0)} \sim \text{Bin}\left(n_{x_1 \dots x_{d_1}}^{(0)}, p_{x_1 \dots x_{d_1}}^{(1|0)}\right).$$

$$P\left(\left|\frac{n_{x_1 \dots x_{d_1}}^{(1)}}{n} - \frac{n_{x_1 \dots x_{d_1}}^{(0)}}{n} p_{x_1 \dots x_{d_1}}^{(1|0)}\right| \geq \varepsilon\right) \leq \frac{n_{x_1 \dots x_{d_1}}^{(0)} p_{x_1 \dots x_{d_1}}^{(1|0)} (1 - p_{x_1 \dots x_{d_1}}^{(1|0)})}{\varepsilon^2 \cdot n^2} \leq \frac{1}{\varepsilon^2 n},$$

so using properties 1 and 4 from Lemma 3, we obtain that $\frac{n_{x_1 \dots x_{d_1}}^{(1)}}{n} \xrightarrow{L} p_{x_1 \dots x_{d_1}} p_{x_1 \dots x_{d_1}}^{(1|0)}$.

Induction Step. Let us consider the stage $i+1$. By the assumption of induction:

$$\frac{n_{x_1 \dots x_{d_i}}^{(i)}}{n} \xrightarrow{L} p_{x_1 \dots x_{d_i}} \prod_{j=1}^i p_{x_1 \dots x_{d_j}}^{(j|j-1)}.$$

After making the selection at the stage i we observe the new features $d_i + 1, \dots, d_{i+1}$, so

$$n_{x_1 \dots x_{d_{i+1}}}^{(i)} | n_{x_1 \dots x_{d_i}}^{(i)} \sim \text{Bin}\left(n_{x_1 \dots x_{d_i}}^{(i)}, p_{x_{d_i+1} \dots x_{d_{i+1}} | x_1 \dots x_{d_i}}\right),$$

where $p_{x_{d_i+1} \dots x_{d_{i+1}} | x_1 \dots x_{d_i}} := P(x_{d_i+1} \dots x_{d_{i+1}} | x_1 \dots x_{d_i})$. Then

$$P\left(\left|\frac{n_{x_1 \dots x_{d_{i+1}}}^{(i)}}{n} - \frac{n_{x_1 \dots x_{d_i}}^{(i)}}{n} p_{x_{d_i+1} \dots x_{d_{i+1}} | x_1 \dots x_{d_i}}\right| \geq \varepsilon\right) \leq \frac{n_{x_1 \dots x_{d_i}}^{(i)} p_{x_{d_i+1} \dots x_{d_{i+1}} | x_1 \dots x_{d_i}} (1 - p_{x_{d_i+1} \dots x_{d_{i+1}} | x_1 \dots x_{d_i}})}{\varepsilon^2 \cdot n^2} \leq \frac{1}{\varepsilon^2 n},$$

hence $\frac{n_{x_1 \dots x_{d_{i+1}}}^{(i)}}{n} \xrightarrow{L} p_{x_1 \dots x_{d_{i+1}}} \prod_{j=1}^i p_{x_1 \dots x_{d_j}}^{(j|j-1)}$.

When we perform the selection at the stage $i + 1$:

$$n_{x_1 \dots x_{d_{i+1}}}^{(i+1)} | n_{x_1 \dots x_{d_{i+1}}}^{(i)} \sim \text{Bin} \left(n_{x_1 \dots x_{d_{i+1}}}^{(i)}, p_{x_1 \dots x_{d_{i+1}}}^{(i+1|i)} \right)$$

then again using Chebyshev-Bienaymé inequality for $\frac{n_{x_1 \dots x_{d_{i+1}}}^{(i+1)}}{n}$:

$$\begin{aligned} P \left(\left| \frac{n_{x_1 \dots x_{d_{i+1}}}^{(i+1)}}{n} - \frac{n_{x_1 \dots x_{d_{i+1}}}^{(i)} \cdot p_{x_1 \dots x_{d_{i+1}}}^{(i+1|i)}}{n} \right| \geq \varepsilon \right) &\leq \\ &\leq \frac{n_{x_1 \dots x_{d_{i+1}}}^{(i)} p_{x_1 \dots x_{d_{i+1}}}^{(i+1|i)} (1 - p_{x_1 \dots x_{d_{i+1}}}^{(i+1|i)})}{\varepsilon^2 \cdot n^2} \leq \frac{1}{\varepsilon^2 n}, \end{aligned}$$

so $\left| \frac{n_{x_1 \dots x_{d_{i+1}}}^{(i+1)}}{n} - \frac{n_{x_1 \dots x_{d_{i+1}}}^{(i)} \cdot p_{x_1 \dots x_{d_{i+1}}}^{(i+1|i)}}{n} \right| \xrightarrow{P} 0$, $n \rightarrow \infty$ and finally:

$$\frac{n_{x_1 \dots x_{d_{i+1}}}^{(i+1)}}{n} \xrightarrow{L} p_{x_1 \dots x_{d_{i+1}}} \prod_{j=1}^{i+1} p_{x_1 \dots x_{d_j}}^{(j|j-1)}.$$

□